Genomics 101 (2013) 204–209

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Methods

Benefit-of-doubt (BOD) scoring: A sequencing-based method for SNP candidate assessment from high to medium read number data sets

Fritz Joachim Sedlazeck [a], Prabhavathi Talloji [b,c], Arndt von Haeseler [a], Andreas Bachmair [b,*]

[a] Center for Integrative Bioinformatics Vienna, University of Vienna, Medical University of Vienna, Dr. Bohr-Gasse 9, A-1030 Vienna, Austria
[b] Department of Biochemistry and Cell Biology, Max F. Perutz Laboratories, Center for Molecular Biology, University of Vienna, Dr. Bohr-Gasse 9, A-1030 Vienna, Austria
[c] Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, D-50829 Cologne, Germany

ARTICLE INFO

Article history:
Received 14 April 2012
Accepted 3 December 2012
Available online 12 December 2012

Keywords:
Next generation sequencing
Read alignment
SNP validation
SNP calling
Sub-genomic library

ABSTRACT

Identification of single nucleotide polymorphisms (SNPs) is a key element in sequence-based genetic analysis. Next generation sequencing offers a cost-effective basis to generate the necessary, large sequence data sets, and bioinformatic methods are being developed to process sequencing machine readouts. We were interested in detection of SNPs in a 350 kb region of an EMS-mutagenized Arabidopsis chromosome 3. The region was selectively analyzed using PCR-generated, overlapping fragments for Solexa sequencing. The ensuing reads provided a high coverage and were processed bioinformatically. In order to assess the SNP candidates obtained with a frequently used alignment program and SNP caller, we developed an additional method that allows the identification of high confidence SNP loci. The method can easily be applied to complete genome sequence data of sufficient coverage.

1. Introduction

Enormous progress has been made regarding the analysis of the genetic basis of phenotypic variation. Classical map (i.e., recombination)-based identification of causative polymorphisms is a routine procedure [1], but has its limits. One of the limits is posed by recombination frequency, which determines how close recombination points can mark the position of interest. Another limit is that some mutant phenotypes are critically dependent on a particular genetic background, so that they are not reliably detected upon out-crossing to a different accession. Detection of a causative polymorphism without the need for outcrossing, as necessary in classical mapping strategies, would therefore significantly extend the range of available methods. In a search for an EMS-induced mutation in *Arabidopsis thaliana*, we faced both of the problems mentioned above, i.e. low recombination frequency in the region of interest, and difficulty of phenotype detection in some mixed Col–*Ler* backgrounds. However, the region of interest was localized to a ca. 350 kb interval on chromosome 3. In order to identify candidate mutations in this region, we applied a sequencing-based approach. Deep sequencing methods have revolutionized genetic analysis [2]. However, data processing is still a subject of optimization and debate. For instance, there is disagreement as to whether high quality reference sequences are necessary to optimally use the benefits of the new methods [3,4]. Furthermore, published deep sequencing strategies make ample use of probabilistic considerations, which are often integrated into data processing algorithms [5,6]. Our source of sequence information, an EMS-mutagenized plant line that was backcrossed several times to the Col-0 progenitor line, is expected to differ at only few positions from the Col-0 wild type line. We found that published methods were not optimally suited for analysis of this data set, resulting in an excess of candidate SNPs. We therefore developed a method to assess candidate SNPs that makes full use of the existing high quality Arabidopsis Col-0 reference sequence, but uses otherwise as few (probabilistic) assumptions as possible. The method is straightforward and transparent in its design, and can be used in a graphic output, or as a score value function.

2. Results

2.1. PCR-based amplification of a region of interest on Arabidopsis chromosome 3

Our region of interest on Arabidopsis chromosome 3, which encompasses predicted genes At3g44400 to At3g44900, consists to approximately 35% of sequence repeats and displays a pronounced suppression of recombination. We therefore considered sequencing-based identification of the mutation of interest. To reduce sequencing costs and to confine the analysis, we generated a library of a subregion of chromosome 3 by PCR amplification.

The amplified sequence covers a region of ca. 347 900 nt of chromosome 3 (Arabidopsis genome sequence version TAIR10). The sequence borders are coordinates 16 042 738 and 16 390 644 of Arabidopsis chromosome 3. We initially designed oligonucleotides with a ca. 20 kb spacing and applied them for PCR with mutant DNA as a template. However, PCR yields with several different “long

* Corresponding author. Fax: +43 1 4277 9748.
E-mail address: andreas.bachmair@univie.ac.at (A. Bachmair).

0888-7543
http://dx.doi.org/10.1016/j.ygeno.2012.12.001

PCR" enzymes were too low, so that we switched to an approximate 11 kb spacing. It was generally straightforward to cover our chromosome 3 region of interest with initially 32 overlapping fragments. In six cases, however, no fragment could be obtained. The average PCR fragment length for the amplified fragments was 11 kb (s.dev. 1.7 kb). In the cases with no initial success, PCR primers were more closely spaced, sometimes in an iterative manner, until satisfactory yields could be obtained. This resulted in an average fragment length of 5 kb for the latter regions (s.dev. 2 kb). Adjacent PCR fragments had an average overlap of 0.45 kb on either side (s.dev. 0.2 kb). Rough measurement of PCR fragment lengths coincided well with expectations based on sequence information. Supplementary Table 1 lists the oligonucleotides used for amplification in their order of appearance on the chromosome, the sizes of amplified fragments and the extent of overlap to the next fragment. All 41 fragments were gel-purified, and DNA concentrations were determined using a nano-drop photometer. The fragments were mixed in equimolar amounts.

2.2. Solexa-based sequencing

5 μg of the equimolar mixture of PCR fragments was sent to a company (GATC of Konstanz, Germany) for Solexa sequencing. In a single run, a total of over 10 million short sequences were obtained, and 9 237 194 were selected by quality control criteria for the alignment. Sequence reads were adjusted to 40 bases to avoid the increased probability of sequencing errors in longer reads.

2.3. Data processing

We aligned the reads obtained from the mutant to the complete sequence of the wild type (WT) reference genome of *A. thaliana* Col-0 (TAIR 10 release). To this end, we applied the program Bowtie 2 [7]. The number of aligned reads depends on whether a read has to align without mismatch, or whether differences (i.e., single nucleotide differences, insertions or deletions) are allowed. At a minimum stringency of 90% (i.e. if the program aligns reads that match at 90% or more of their sequence), nearly all of the reads (98%) can be aligned. The average number of aligned reads per position is 996.12 (median 686). With increasing percentage of matching nucleotides, less and less reads can be aligned. If only exactly matching reads are admissible, the average number of aligned reads per position drops to 846.95 (median 587). By plotting the fraction of aligned reads against the number of differences maximally allowed in the alignment process (Fig. 1), we observed that the increase from perfectly matching reads to reads with up to four differences is not uniform. When comparing 90% identity and 100% identity (red vs. green arrow of Fig. 1), the difference comprises almost 15% of all reads. Of particular interest to us was that most of the difference is actually between ca. 97% and 100% read identity. A read of length 40 bases with 97% identity can at most have one differing position, indicating that this class of reads is informative regarding the detection of single nucleotide polymorphisms (SNPs). Further bioinformatic processing is expected to extract information regarding SNP candidate positions. In our sample, we compared a mutagenized sequence to its progenitor sequence, and therefore we did not expect a large number of SNPs. In fact, we expected that most of the reads alignable with a similarity of 90%, but not alignable with 100% match, are due to errors in the sample preparation and sequencing procedure, and cannot be attributed to true differences in the sequence.

SAMtools [8] was applied to obtain a list of candidate SNP positions. 32 positions were listed by the program as potential SNPs. 14 of them mapped to the region of interest. Calling of candidate SNP positions in other regions of the Arabidopsis genome may be a consequence of the fact that our region of interest is particularly rich in sequence repeats, leading to erroneous alignment of some reads to

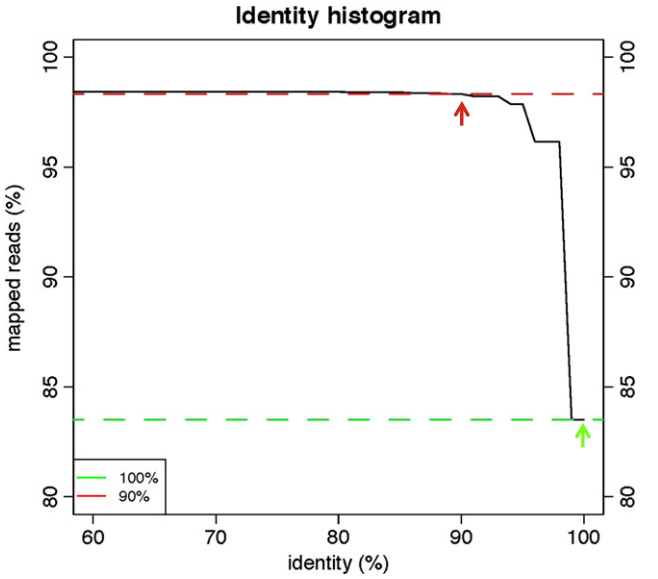


Fig. 1. Fraction of reads aligned as a function of their matching bases. The number of reads aligned to the genomic region of interest is a function of the chosen identity threshold. At a threshold of 90% identity or lower, the number of reads that can be aligned to the reference sequence is almost constant. If a higher threshold of identity is chosen, more and more reads cannot be aligned. The steepest decline occurs between 97% and 100% identity and reflects a difference between one nonmatching position per read and exactly matching reads (length 40 bases). Black line, fraction of aligned reads; red arrow, 90% identity threshold; red dotted line, percentage of mapped reads at the 90% threshold; green arrow, 100% identity threshold; green dotted line, percentage of mapped reads at the 100% threshold.

other regions of the Arabidopsis genome. However, 14 EMS-induced mutations in the 350 kb region of interest seemed still too high, because this mutation density across the complete genome in the original EMS-treated plant line would seriously compromise plant growth, which was not observed. We therefore sought to establish an additional filtering process that is better able to pinpoint true SNP positions in our sequence data set.

To that end, we determined the position wise coverage by reads aligning along the WT genome, for regions around SNP candidate positions as obtained from SAMtools. The coverages were computed for two categories of reads. Category 1 (C-90) comprises all reads that show at least 90% identity to the WT, category 2 (C-100) gathers all reads that are 100% identical to the WT. Thus, for each position in the WT genome we obtained two counts, the number of C-100 reads and the number of C-90 reads. Obviously the latter is always larger than or equal to the former and, as a further characteristic value, we computed the position-wise and the average ratios C-90 reads divided by C-100 reads (C-90/C-100 ratio). The average ratio C-90/C-100 in the region of interest is 1.157 (median: 1.160). If most of the reads at a given position carry the base of the WT, the ratio should be close to the average. In contrast, if a particular position differs from the WT base, the number of reads within the tolerated 10% mismatch level would not differ significantly from local neighbors. However, only very few reads with perfect fit (0% mismatches) can be expected at such positions, as generation of reads with WT sequence at mutant positions requires actually a sequencing error.

Fig. 2 displays the coverage of C-90 (red line) and C-100 (green line) for the 14 potential mutation spots. Similar graphs for the other 18 SNP candidate positions (that lie outside the sequenced region) are shown in Supplementary Fig. 1. We obtain essentially three patterns. One pattern consists of an irregularly shaped curve with sharp drops and/or increases for both the C-90 and C-100 coverage. Interestingly, when the graphic display is used for SAMtool SNP

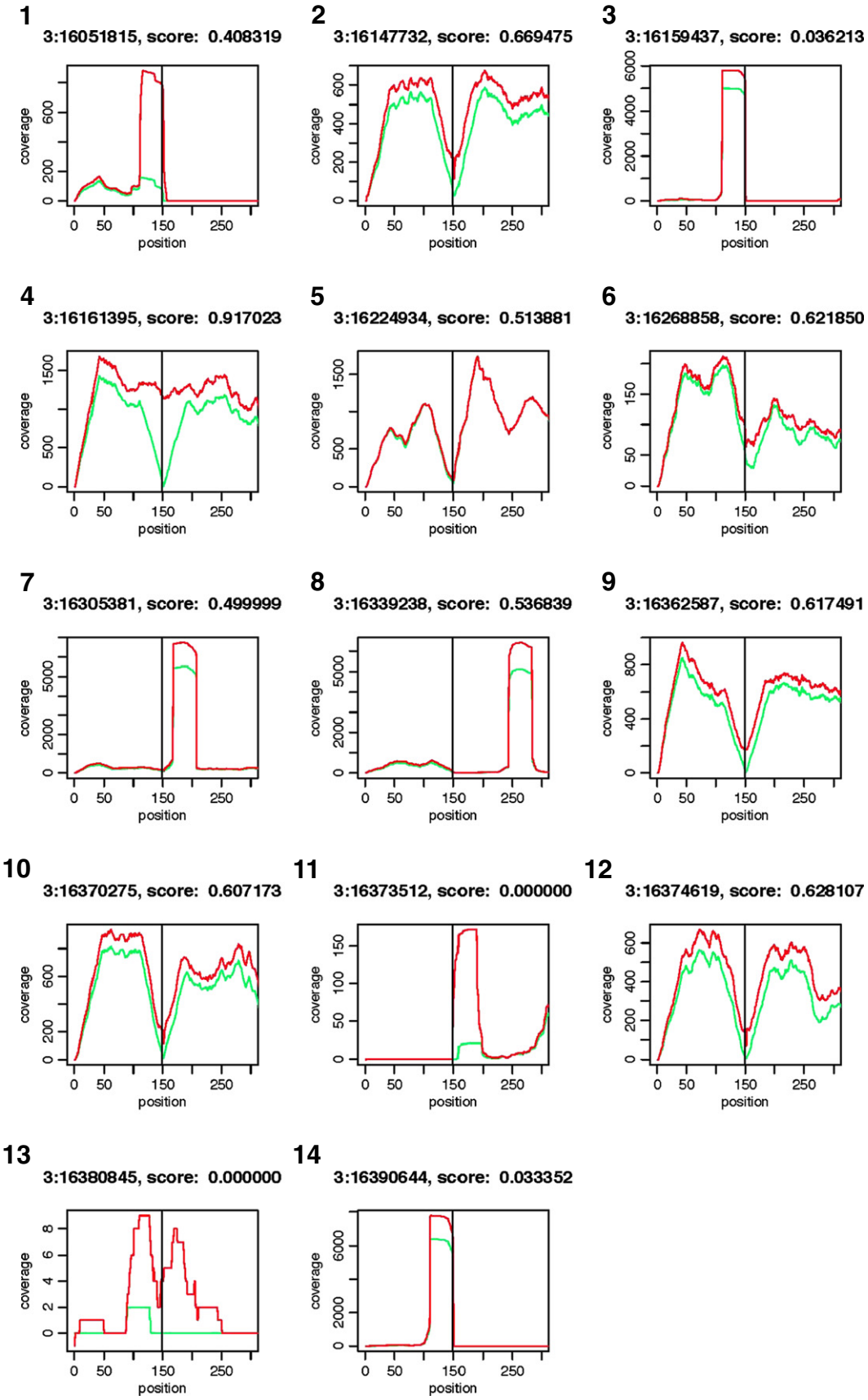


Fig. 2. C-90 and C-100 read frequencies at predicted SNP positions, numbered according to occurrence on the chromosome. The number of reads was plotted on the y-axis for the C-90 (red lines) and the C-100 (green lines) set, respectively. The x-axis depicts a window of 150 bases before and after the base of a predicted SNP, which is delineated by a black vertical line. Chromosome number, position on the chromosome and BOD score value are written above each graph.

candidate positions that lie outside of the sequence region of interest, all of them display this pattern (Supplementary Fig. 1). A second pattern consists of a more gradual drop and subsequent rise in coverage for both the C-90 set of reads and the C-100 set. This pattern is represented by SNP candidates 2, 9, 10, and 12 of Fig. 2. In the third pattern, represented by SNP candidate 4 of Fig. 2, the C-100 set displays a gradual drop and subsequent rise in coverage, whereas the C-90 set stays more or less constant across this region. It was our expectation that this latter pattern should be most characteristic of true SNPs.

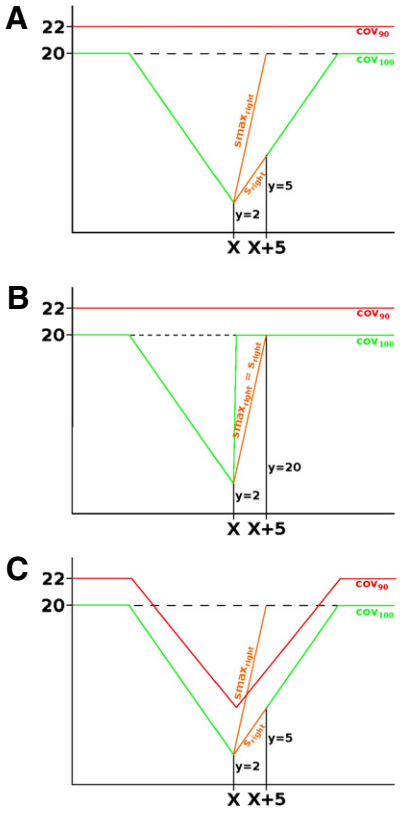
In addition to a graphic output, we wanted to establish a scoring function with output value between 0 (poor score) and 1 (highest possible score). The score was designed to reflect the presence or absence of a V-shaped C-100 read distribution and the constant C-90 coverage. The formula for the BOD score is depicted in Fig. 3. It puts half of its weight on the V-shape of the C-100 data set. The other half is put on the coverage by the C-90 data set across the same region. Interestingly, the class of SNP candidates with V-shape in both the C-100 and the C-90 data set still score high with this formula, giving a value above 0.6, whereas the SNP candidates with irregular pattern give a value below 0.6. Score values for all graphically depicted SNP candidate positions are listed above the graphs (Fig. 2 and Supplementary Fig. 1).

2.4. Validation of primary results

Predicted SNP candidates that lie outside the sequenced region, as shown in Supplementary Fig. 1, were considered as false predictions without further experimentation. To experimentally validate the predicted mutant spots lying within the region of interest, we applied Sanger sequencing on PCR fragments amplified from the six representative SNP candidates numbered 1, 4, 8, 9, 11 and 12 in Fig. 2, comprising the range of graphic patterns and BOD scores. Results are listed in Table 1, in the order of decreasing BOD score. We found that only one of the SNP candidates with irregular pattern turned out to contain a deviation from the reference sequence. In contrast, the pattern with V-shaped C-100 read set, but relatively constant C-90 read set, is apparently indicative of a true SNP. The pattern with V-shape of both C-100 and C-90 was also of high interest. We expected that it might on the one hand be indicative of an unusual distribution of aligned sequence reads, resulting in a false call. On the other hand, it may also be that at the position of the potential SNP, the sequence deviates by more than one mismatching nucleotide, so that even C-90 (tolerating a difference of four nucleotide changes) is not tolerant enough for alignment. The two patterns of this kind validated in Table 1 turned out to fall into this latter class. We find a one base insertion, and a nucleotide change combined with a short insertion, respectively. From these results, we propose that the method introduced offers a promising accuracy prediction for potential deviation candidates.

3. Discussion

In this work, we present a simple method to evaluate predicted deviations from the reference sequence (SNPs and small indels) in the model plant *A. thaliana*. The use of sequencing for mutant identification after crossing of two different accessions was shown recently [9–13]. We found that established SNP callers were not optimally



D

$$slope(y_1, y_2) = \frac{y_2 - y_1}{5},$$

$$s_{right} = slope(cov_{100}[pos], cov_{100}[pos + 5]),$$

$$s_{left} = slope(cov_{100}[pos], cov_{100}[pos - 5]),$$

$$smax_{right} = slope(cov_{100}[pos], \max_{pos \le i \le pos + |read|}(cov_{100}[i])),$$

$$smax_{left} = slope(cov_{100}[pos], \max_{pos - |read| \le i \le pos}(cov_{100}[i])),$$

$$s_{comb} = 1 - ((\frac{s_{left}}{smax_{left}} + \frac{s_{right}}{smax_{right}})/2),$$

$$cov = cov_{90}[pos] / \max_{pos - |read| \le i \le pos + |read|}(cov_{90}[i]),$$

$$score = \frac{s_{comb} + cov}{2}$$

Fig. 3. Prototypic read frequency diagrams and BOD score formula. Diagrams such as depicted in Fig. 2 were idealized to pinpoint parameters measured to obtain the BOD score value of a particular SNP candidate. A, idealized diagram of a high scoring SNP candidate (BOD score close to 1). B, a one-sided steep drop in C-100 read coverage already reduces the score (BOD score ca. 0.7). C, idealized diagram of a position with either fluctuating coverage, or with a sequence deviation exceeding the mismatch tolerance of the tolerant alignment set C-90 (BOD score ca. 0.6). D, formula. The formula measures the read number of the locally aligned C-100 reads by comparison of the SNP candidate position with a position at a 5 base distance for either direction. This slope is put in relation to the maximally possible slope, given by the average number of reads in the vicinity of the SNP candidate. The relative difference in read number (SNP position versus local average) is also determined for the tolerant alignment data set C-90. The final score lies between 0 and 1 and is high if the C-100 read numbers display a not-to-steep V shape at the position of interest, whereas the C-90 read number does not decrease at the same position.

Table 1
Sequence verification of selected candidate single nucleotide polymorphism regions[a].

BOD score	Coordinate on chromosome 3	Sequence context[b]	Confirmation process result	Annotation
0.917023	16 161 395	TGGAAA**G**AGGAAA TGGAAA**A**AGGAAA	Confirmed	Between annotated ORFs At3g44580 and At3g44590
0.628107	16 374 619	CCGAAG TGACAAC CCGAAG**G**TGACAAC	Confirmed[c]	Before the stop codon of ORF At3g44850
0.617491	16 362 587	CTACTA **A** TCGCCA CTACTA**GCT**TCGCCA	Confirmed[c]	Within ORF At3g44820
0.536839	16 339 238	TTCAGA**A**GCTTGA TTCAGA**C**GCTTGA	Confirmed[d]	Within ORF At3g44796
0.408319	16 051 815	AGCTTC**A**GGGTTT AGCTTC**A**GGGTTT	Not confirmed	Between annotated ORFs At3g44400 and At3g44410
0.000000	16 373 512	GAGGTC**T**ATATTG GAGGTC**T**ATATTG	Not confirmed	Within ORF At3g44840

[a] Selected genomic regions were amplified by PCR from mutated and progenitor line and subjected to conventional sequencing, with the results as listed.
[b] The reference sequence context is written at the top, the sequence of the EMS-treated plant line as determined by confirmatory (Sanger) sequencing is written below. The nucleotides in question are shown in bold.
[c] The sequence change includes a small insertion/deletion compared to the reference sequence. However, the altered sequence of the mutant line was also present in the non-mutagenized progenitor line, suggesting that the nucleotide change cannot cause phenotypic differences between progenitor and mutant plant lines.
[d] SNP was confirmed, but the differing nucleotide is present both in the mutagenized line and in its progenitor, suggesting that the nucleotide change cannot cause phenotypic differences between progenitor and mutant plant lines.

suited for our data set, resulting in a high fraction of false positive candidate SNP predictions. We present a filter that can be used to further evaluate the output of popular and upcoming SNP callers.

The method introduced in this work profits from increases in sequence coverage and exploits the availability of a high quality reference sequence. Expressed in non-mathematical terms, we asked for each position the question, whether it is reasonable to assume that the position contains the base of the reference sequence (whereas most SNP callers ask the question whether there is a chance that the position contains a SNP). There are positions in the sequence where exceptionally few reads with perfect match are aligning, compared to the number of aligned reads when differences are tolerated. These positions are prime candidates for the presence of a SNP. Because this treatment of data puts excessive emphasis on the reference sequence, giving the presence of the reference nucleotide at each investigated position the "benefit of doubt", we called this treatment of data benefit-of-doubt (BOD) scoring. The prime source of information is therefore reads that show perfect alignment (called C-100), and their absence at a particular position is critically noticed. However, the second data set, reads that align with a certain error margin (in our example, C-90), also play an important part. They are used as a control of the coverage distribution within a small region surrounding the SNP position. The importance of this type of control is particularly obvious at positions that are "false positives" (as determined by verification of candidate positions in our sequence read set). The largest set of false positive SNP calls in our data set is positions where reads were wrongly aligned, to regions outside the sequenced area (graphs depicted in Supplemental Fig. 1). None of these positions fulfills the criteria defined as optimal by BOD scoring. For a selected set of SNP candidate positions in the sequenced region, Sanger sequencing was chosen for verification (Table 1). One verified SNP showed the only case where the V-shaped distribution of C-100 reads was observed in combination with the relatively constant coverage distribution of the C-90 reads. This is the optimal case, and in addition to identification by a graph (Fig. 2), the BOD scoring function introduced in this work posits a high value (0.92) for this type of read distribution. We also re-sequenced two positions (second and third entries of Table 1, BOD scores 0.63 and 0.62, respectively, graph shown in Fig. 2) where the deviation in sequence allowed neither local alignment of C-100 reads, nor of reads from the C-90 data set, because of a small insertion. Thus, it may also be worth considering positions where both the C-90 and the C-100 set of reads display a V-shape, or one may consider a more tolerant error limit for the not perfectly aligning set (in particular, allowing short gaps). In contrast, both C-100 and C-90 read numbers apparently fluctuate significantly and non-symmetrically at most positions that SAMtools suggests as SNP candidates, but where the BOD scoring function assigns a low score. As an example, most wrong SNP candidates shown in Supplemental Fig. 1, and, likewise, the two not confirmed SNP candidates listed in Table 1, have a BOD score below 0.45.

While it is clear that the randomness of sequence generation results in fluctuations in coverage, graphic displays as shown in Fig. 2 and Supplementary Fig. 1 indicate that at some positions, the transition from higher to lower coverage can be very abrupt. If, at these positions, there is also a higher probability of a sequencing error, this might lead to an erroneous call of an SNP. The V-shape output is an efficient buffer against such false positive SNP calls, because the V symmetry requires a gradual coverage change and roughly constant coverage on both sides of the potential SNP, and thereby eliminates positions with abruptly fluctuating coverage.

Regarding extension of this method to other data sets, we propose that the method is also applicable to whole genome sequence data sets, as the difference is largely dealt with by the alignment and SNP caller programs. Another obvious question pertains to the presence of heterozygous SNPs in some data sets. We are optimistic that both the graphic display, and the BOD scoring function are a big help in pinpointing such candidate positions, as well, because the V shape is also identifiable if the bottom of the V does not reach zero y value, and the BOD score scalar value measures steepness of decline, not absolute value at the minimum. However, it is clear that these issues can only be clarified by application on real data sets.

In the method presented, reads from a mutated genome are aligned to the corresponding sequence of the WT, here Arabidopsis accession Col-0, the data are analyzed by an SNP caller, and the output is further evaluated by taking local read coverage and its changes into account. In conclusion, we introduce a sequencing-based method to select and score candidate positions for SNPs and mutations in plants with high quality reference genome sequence such as Arabidopsis accession Col-0.

4. Materials and methods

4.1. PCR reactions to generate a sub-genomic fragment pool

Arabidopsis DNA was prepared as described [14], using a small-scale preparation method. Amplification of PCR fragments with the primers listed in Supplementary Table 1 was carried out using Finnzymes Phusion Hot Start High Fidelity DNA Polymerase

(Thermo Fisher Scientific) as recommended by the manufacturer. PCR conditions were as follows: initial denaturation 95 °C for 7 min, followed by 35 cycles of 95 °C for 30 s, 50 °C for 30 s, 68 °C for 12 min, followed by 68 °C for 10 min as a last elongation step before sample cooling. In a few cases, better results were obtained using LA Taq (TaKaRa BIO) as recommended by the manufacturer. PCR conditions in the latter case were an initial denaturation step of 95 °C for 5 min, followed by 34 cycles of 95 °C for 30 s and 68 °C for 8.5 min. A final elongation step (68 °C, 10 min) followed before cooling of the sample. PCR fragments were gel-purified (Wizard SV Gel and PCR Clean-Up System; Promega), and adjusted to 100 pmol/μl in water.

4.2. Analysis of SNP candidates

After using Bowtie 2 [7] for alignment of reads to the reference *A. thaliana* sequence (release TAIR10), SAMtools [8] was used to obtain a list of SNP candidate positions. For further evaluation of candidate positions, we computed for each mapped read in a candidate region the percent identity, that is, the number of identical nucleotides (matches) in the alignment to the reference, divided by the alignment length. Secondly, we counted for each region of a predicted SNP position the number of mapped reads that show a minimal percent identity of either 90%, or of 100%. We also computed the position-wise ratio C-90/C-100 for the complete sequenced region. If the read count of C-100 was equal to 0 at a position, then the C-90 count was used directly (i.e., without division) to avoid division by 0.

4.3. BOD score function and generation of graphic output

The computation of the BOD score consists of two parts, as shown in Fig. 3. Firstly, we computed the slope of coverage between the candidate SNP position (x) and the position x′, which was by default set 5 bp apart. The value of 5 was chosen as a small, but sizable fraction of the read length (12.5% of the read length of 40 bases) In addition, we computed the maximum observed coverage (max_cov) in the region from x to x + read length. The ratio between the slope of the SNP position (x,y1) and the point (x′,y2) and the slope of the SNP position and the point (x′,max_cov) is computed. Here, y indicates the coverage of reads that are 100% identical to the reference sequence at the position. The calculation is carried out for the region left (−5 bp distance) and right (+5 bp distance) of the candidate SNP position. We used the average ratio from both regions to determine the similarity to a V shape. Secondly, we computed the ratio of the C-90 coverage between the candidate SNP position and a surrounding region that equals twice the read length and overlaps the SNP position. This ratio indicates whether there is a general drop of coverage near this position, which might lead to a V-shape, or whether the coverage is more or less even. The BOD score is the average of the value indicating the V-shape of C-100 reads, and the value indicating the constant coverage by C-90 reads. For graphs, we generated a tab separated file of the coverage of the particular regions. These values were read into R and plotted. Computational tools for BOD score are available under the following URL: http://cibiv.github.com/BODscore.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2012.12.001.

Acknowledgments

We want to thank Marcus Garzón for initial mapping experiments and Bulut Hamali for contributions to SNP validation. Work in A.B.'s lab was supported by the German Research Foundation DFG (grant BA1158/3-1), by the Max Planck Society, and by the Austrian Science Foundation FWF (grant P 21215-B12). Work in A.v.H.'s lab was supported by DFG SPP1174 (grant HA1628-9-2) and the Wiener Wissenschafts- und Technologie-Fonds (WWTF).

References

[1] X. Hou, L. Li, Z. Peng, B. Wei, S. Tang, M. Ding, J. Liu, et al., A platform of high-density INDEL/CAPS markers for map-based cloning in Arabidopsis, Plant J. 63 (2010) 880–888.
[2] J.P. Hamilton, C.R. Buell, Advances in plant genome sequencing, Plant J. 70 (2012) 177–190.
[3] C. Feuillet, J.E. Leach, J. Rogers, P.S. Schnable, K. Eversole, Crop genome sequencing: lessons and rationales, Trends Plant Sci. 16 (2011) 77–88.
[4] K. Schneeberger, D. Weigel, Fast-forward genetics enabled by new sequencing technologies, Trends Plant Sci. 16 (2011) 282–288.
[5] R. Nielsen, J.S. Paul, A. Albrechtsen, Y.S. Song, Genotype and SNP calling from next-generation sequencing data, Nat. Rev. Genet. 12 (2011) 443–451.
[6] S. Ossowski, K. Schneeberger, R.M. Clark, C. Lanz, N. Warthmann, D. Weigel, Sequencing of natural strains of *Arabidopsis thaliana* with short reads, Genome Res. 18 (2008) 2024–2033.
[7] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (2012) 357–359.
[8] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 genome project data processing subgroup, the sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.
[9] R.S. Austin, D. Vidaurre, G. Stamatiou, R. Breit, N.J. Provart, D. Bonetta, et al., Next-generation mapping of Arabidopsis genes, Plant J. 67 (2011) 715–725.
[10] R.A. Laitinen, K. Schneeberger, N.S. Jelly, S. Ossowski, D. Weigel, Identification of a spontaneous frame shift mutation in a nonreference Arabidopsis accession using whole genome sequencing, Plant Physiol. 153 (2010) 652–654.
[11] N. Uchida, T. Sakamoto, T. Kurata, M. Tasaka, Identification of EMS-induced causal mutations in a non-reference *Arabidopsis thaliana* accession by whole genome sequencing, Plant Cell Physiol. 52 (2011) 716–722.
[12] K.-H. Lin, M. McCormack, J. Sheen, Targeted parallel sequencing of large genetically-defined regions for identifying mutations in Arabidopsis, Plant Methods 8 (2012) 12.
[13] K. Schneeberger, S. Ossowski, C. Lanz, T. Juul, A.H. Petersen, K. Lehmann Nielsen, et al., SHOREmap: simultaneous mapping and mutation identification by deep sequencing, Nat. Methods 6 (2009) 550–551.
[14] A. Tramontano, A. Donath, S.H. Bernhart, K. Reiche, G. Böhmdorfer, P.F. Stadler, A. Bachmair, Deletion analysis of the 3′ long terminal repeat sequence of plant retrotransposon Tto1 identifies 125 base pairs redundancy as sufficient for first strand transfer, Virology 412 (2011) 75–82.